“大数据”专题

doi: 10.3969/j.issn.1673-5692.2013.01.005

大数据时代

(《中国电子科学研究院学报》编辑部,北京 100846)

摘 要: 分析了大数据的产生,归纳了大数据的“大”这个特点,可以用多样化、海量、快速、灵活和复杂“4V+1C”来概括;从大数据的存储问题、处理问题、部门之间的信息是否融合几个方面论述了大数据对通信、医药、天文学、宇宙学、材料科学、气象学等领域产生影响;最后,提出了几条应对措施。

关键词: 大数据;信息化;影响

中图分类号: TP391 **文献标识码:** A **文章编号:** 1673-5692(2013)01-027-05

The Big Data Age

(Editorial Office of《Journal of China Academy of Electronics and Information Technology》,Beijing 100846,China)

Abstract: Analyzed the generation of big data, summarized the characteristic “big” of big data, it can be summarized by diversification, mass, fast, agile and complicated “4 v + 1 c”. From the big data storage, processing, information fusion between departments, big data impact is discussed in respects of communication, medicine, astronomy, cosmology, materials science, meteorology, etc. Finally, put forward several countermeasures.

Key words: big data; Informatization; influence

0 引 言

“大数据”在 2011 年一路走红,在 2012 年更加闪耀,成为业界当之无愧的焦点。尤其是随着新型 SNS 网络的发展、视频流量的猛增及图片分享需求的涌现,大数据在肥沃的土壤中更加迅速的成长。Forrester Research 分析师表示,大数据意味着规模极大的分析量,意味着高速处理批比特(千万亿字节)的结构化数据及非结构化数据的能力。对于企业机构来讲,大数据是一把双刃剑。一方面,机构能够从更多的渠道获得更加丰富的关于用户的信息;另一方面,当前的数据分析能力却对大规模的非结构性数据国不从心。从理论上而言,如果能够从冗杂的大数据中剥丝抽茧,识别出最有价值的信息并进行分析处理,那么将会更精准准备地把握市场趋势。

1 “大数据”的产生

“大数据”是一个术语,是一个带有文化基因和营销理念的词汇,但同时也反映了科技领域中正在发展中的趋势,这种趋势为理解这个世界和作出决策的新方法开启了一扇大门。“大数据”的出现不是一个偶然的事情,它是在信息化、网络化高度发达的今天,在这个数据整天飞的时代所必须要经历的过程。这个现象的出现同时又给网络安全及维护,信息攻击及防御带来了新的问题和挑战。

那么到底什么是“大数据”呢?

维基百科上说:大数据指的是“网络公司日常运营所生成和积累用户网络行为”,数据增长如此之快,以至于难以使用现有的数据库管理工具来驾驭,困难存在于数据的获取、存储、探索、共享、分析和可视化等方面。数据量的增长到现在,已经不是以我们所熟

收稿日期: 2013-01-21 修订日期: 2013-01-31

知的多少 G 和多少 T 来描述了,而是以 P(1 千 T),E(1 百万 T)或 Z(10 亿 T)为计量单位。百度对此给予了更形象的描述。光是其新首页导航每天就要从超过 1.5PB 的数据中进行挖掘,这些数据如果打印出来将超过 5 千亿张 A4 纸,摞起来会超过 4 万公里高,接近地球同步卫星轨道长度,平铺可以铺满海南岛。而 2020 年新增的数字信息成长幅度将是 2009 年的近 45 倍。如今,只需要两天就能创造出自文明诞生以来到 2003 年所产生的数据总量。

这些数据的规模、形式超出了传统数据处理方法所能捕获、管理和处理的能力。人类的这种能力是不断进步的,所以,大数据的数据量是一个不断变化的目标。美国地理空间情报基金会(USGIF)关于的一个大数据的情况讨论会中提到人类现在处理数据能力的增长速度如今跟不上数据量的增长速度,所以,在数据面前,处理能力总是有很大的空缺。过去做信息处理的方法应该要尽快做出调整,掌握大数据的处理能力,会使得在之后的信息处理各个领域掌握主动权。

另外,大数据,不仅仅是指大量的数据,也不是仅仅指数据的指数增长速度,它是对需要对当前架构需要做出调整的数据进行在理解上的新的方式和理念。对数据整合算法,数据结构理解使用上的新的方法的研究势在必行。

2 "大数据"的特点

CIO 时代网(www.ciotimes.com)总结出,"大数据"不仅有"大"这个特点,还有很多其他的特色。总体而言,可以用"4V +1C"来概括。

(1) Variety(多样化)

大数据一般包括以事务为代表的结构化数据、以网页为代表的半结构化数据和以视频和语音信息为代表的非结构化等多类数据,并且它们的处理和分析方式区别很大。

(2) Volume(海量)

通过各种智能设备产生了大量的数据,PB 级别可谓是常态,笔者接触的一些客户每天处理的数据量都在几十 GB、几百 GB 左右,估计国内大型互联网企业每天的数据量已经接近 TB 级别。

(3) Velocity(快速)

大数据要求快速处理,因为有些数据存在时效性。比如电商的数据,假如今天数据的分析结果要等到明天才能得到,那么将会使电商很难做类似补货这样的决策,从而导致这些数据失去了分析的意义。

(4) Vitality(灵活)

在互联网时代,和以往相比,企业的业务需求更新的频率加快了很多,那么相关大数据的分析和处理模型必须快速地适应新的业务需求。

(5) Complexity(复杂)

虽然传统的商务智能(BI)已经很复杂了,但是由于前面 4 个 V 的存在,使得针对大数据的处理和分析更艰巨,并且过去那套基于关系型数据库的 BI 开始有点不合时宜了,同时也需要根据不同的业务场景,采取不同的处理方式和工具。

以上新时代下"大数据"的特点决定它肯定会对当今信息时代的数据处理产生很大的影响。

3 "大数据"对当今信息时代产生的影响

随着数据生成和采集的指数增长,不管是由于下一代望远镜,高通道的科学试验,还是千万亿次的科学计算,高分辨率的传感器,以及更加错综复杂的网络环境,大数据的出现在科学道路上是一个让人兴奋的时代。由于这些高科技的出现,它将在未来十年内对通信、医药、天文学、宇宙学、材料科学和气象学等领域造成更加显著的影响。同时,将会发现,在以前处理低数据量的时候所用的方法和技术可能在当前大数据的条件下,不会再起到应该有的效果。在高通量的数据传递速率的条件下,需要更高更先进的技术去对数据进行采样描述分析,这对新技术、新设备的研究开发提出了更高的要求。

大数据的存储问题。随着越来越多的视频、影像、出版、分析和虚拟化等内容的文件越来越多,单个文件的大小和容量日益增加,在这样的情况下,如何对这些"大"数据文件进行更加有效合理的管理成为企业用户面临的一个问题。与管理传统的非"大"数据文件不同,管理这些"大"数据文件面临以下几个问题:首先是高性能共享的问题,由于数据容量大,这就对数据共享的性能提出了挑战,传统的"小"数据的存储解决方案显然不会得到好的性能。其次是文件管理和保护的问题,由于文件个头变大,对它进行分级、归档、备份和保护等都将对整个数据传输网络的性能提出严峻的挑战。最后,是重复数据的问题,大量重复的"大"数据文件肯定会占用更多的存储资源。

大数据的处理问题。过去的科学研究第三范式

就已经需要用计算机来处理大型的数据运算和模拟。而如今,这些研究正在被大量的数据淹没。数字信息从各种各样的传感器、工具和模拟实验那里源源不断地涌来,令组织能力、分析能力和储存信息的能力捉襟见肘。科学家将会在天文观测、气象监测、生物基因、物理仿真等数据密集型科学研究中遭遇大数据这一问题。

大数据时代的多种数据处理的可能性不会只限于我们对传统的交互、输入、输出、搜索的理解。Petabyte 数据量级别的数据提供了我们去思考数据在信息中的新角色和数据间关系的机会。

在管理与政策上,大数据时代下面临的问题包括企业或政府机构跨部门的信息是否能融合,而且更为重要的是个人隐私等信息安全问题能否得到解决。其中最为迫切需要解决的就是安全问题。这里所说的安全不同于以往的信息安全问题,而是一种新的安全观。这种新安全观需要在大数据的利用时找到开放和保护的平衡。例如涉及个人隐私的数据,既要能够深入挖掘其中给人类带来利益的智慧部分,又要充分保护隐私数据不被滥用,损害到个体的利益。

另一个挑战则是大数据人才的培养。仅美国就面临 14 万至 19 万分析和管理人才缺口,以及 150 万具备理解和基于大数据研究做出决策的经理和分析师人才缺口。因而,能让大数据对商业更有利和更有价值的分析和管理人才还比较有限。

在新形势下,世界各地出现的数据危机逐渐显现出来,据国外媒体报道,美国联邦执法部门和情报机构在网上发布的信息征集启事显示,美国政府正在寻找一款能够分析社交媒体海量数据,并预测未来恐怖主义袭击和国外暴乱等重大事件的软件。FBI 透露它希望借助数据工具来扫描和分析整个社交媒体中的庞大数据。美国国防部和情报局总监办公室(Office of the Director of National Intelligence)也已向私有企业求谋良策,希望利用社交媒体上人们每日共享的数十亿条帖子来识别可能会发生的突发事件,例如恐怖主义威胁和骚乱活动。

在情报界,分析公众信息并不是什么新鲜事。例如,在冷战时期,美国中央情报局(CIA)的特工人员就经常阅读俄罗斯新闻报纸,拦截他们的电视和广播节目,企图推断苏联领导人正在想什么。在过去几年中,社交媒体的崛起极大地改变了公众信息的数量和类别。Twitter CEO 迪克 - 科斯特罗(Dick Costolo)在最近一次会议中声称,该微博网站的用户平均每三天发布 10 亿条消息。CIA 前分析师罗斯 - 斯塔普勒顿 - 格雷(Ross Stapleton-Gray)说,“现在是收集情报的黄金时代,因为所有人都在自觉自愿地表达他们是谁”。在 20 世纪 90 年代初,格雷供职于 CIA 总监办公室。他现在是一名技术顾问,为公司提供安全、监控和隐私等方面的建议。格雷声称,美国情报机构早期收集互联网信息的努力,遭到了一些元老级人物的阻扰,他们坚信机密信息比任何人均能够获取的互联网信息更有价值。但是,这些机构寻找最佳社交媒体分析工具的做法表明,这种阻力已经大大减弱了。

美国情报局总监办公室下属的研究部门致力寻找的软件系统,将会融合网络研究到维基百科编辑到流量监控等各种功能,而且将能够预测未来可能发生的重大事件,包括从经济混乱到瘟疫爆发。美国国防部寻找的工具将跟踪社交媒体,监测那些可能影响作战士兵情绪的信息的传播,并让军方在社交网络上执行“有效的网络作战方案”,打击各种敌对活动。美国情报局总监办公室和国防部声称,他们不会在美联社要求的期限内回答有关这项提议的具体问题。

针对这些暴露出来的新型问题,必须要采取相应的应对措施来维持一个良好的社会秩序、科研环境、网络环境。

4 如何适应“大数据”时代

如果问计算科学的专家,在今天什么将会使他们在自己领域有更大的进步,大多数的人都会说是更大的磁盘空间和更快的 CPU 速度。但是如今,新兴的 petabytes 级别的数据量从根本上改面了他们的认识,新的工具(电脑硬件和软件),新兴技术(算法和统计规律)和科学计算周期本身都会同时加快进步的速度。

(1)加强领域合作。在科学研究上,在高通量的数据流不断涌出,多种数据形式并存的情况下,要分清数据是结构化的还是非结构化的,从而进行针对性分析。同时,要实时的决定哪些数据应当被保留,而哪些数据是要被舍弃的。我们必须要确定访问的目标资源和资源的组织方式的最佳组合。这种数据处理方式的产生和对整体的数据问题的分析,需要计算机科学技术科学家和目标专家的密切的合作。

(2)开发数据密集型计算方法。在信息量呈指数级增长之时,必须重新考虑数据密集型科学的一整套方法。图灵奖得主、已故科学家吉姆·格雷(Jim Gray)针对这种情况提出了“第四范式”(the

fourth paradigm) 。吉姆认为: 人类需要用强大的新工具去分析、呈现、挖掘和处理科学数据。要解决我们面临的某些最棘手的全球性挑战,它们可能是唯一具有系统性的方法。另一方面,科学研究的第四范式发展了一种"众包"研究模式,例如海洋研究项目来说,如今对海洋的观测会产生海量的信息,这些信息如果得不到合理的组织和存储,后续研究就无法开展。因此,为了确保任何一个研究机构不会因此不堪重负,他们让世界各地的科学家、学生和感兴趣的民众都可以访问这些数据。此外,谷歌也在运用数据处理技术解决科学和社会问题。如由其发起的地球引擎(Earth Engine) 项目: 使用卫星图像和卫星分析技术,对全球森林沙漠化进行跟踪; 登革热和流感趋势(Dengue & Flu Trends) 项目: 通过汇总 Google 搜索数据,用以估计近乎实时的疾病活动; 危机响应(Crisis Response) 项目: 提供重要信息和开发工具,用于支持抗灾救危; REC 项目: 致力于开发寻找比煤廉价且能达到公用事业规模的可再生能源的工具。谷歌公司的这一系列项目将大数据的分析淋漓精致地用到了科学研究中,为科学创新提供了源动力。

美国能源部已经是而且将继续会是在高性能数据计算方面的先导者之一。要在一些代表性的模拟试验中获取最好的结果,他们需要具备产生和管理大宗数据的能力,此外,还需要相关的工具来从数据中提取有用的信息进行分析。在能源部的数据密集型科学中,面对案例研究和未来挑战,James P. Ahrens 和他的合作者梳理出一套在处理这些数据时会遇到的一些共同的挑战。这些挑战包括网络和分析的基础设施,从大规模气象学及宇宙学模拟得到的数据,X - 射线观测站的数据,和从能源部用户设施的中子散射数据,这些用户设施包括阿贡国家实验室的高级光子源和美国国家散裂中子源。

Randal E. Bryant 在他的一篇文章"可扩展的数据密集型科学计算应用"中指出了在数据密集型的科学计算中可扩展性的重要性。不管是在管理数据还是执行大量的数据计算的时候都应如此。Bryant 提出将来的数据密集型的科学计算系统将会明显不同于较为传统的 HPC 系统。HPC 系统是在当前多数设备还在使用的计算装置。在"数据密集型科学计算"中,Alexander S. Szalay 关注到了天文学界(不止此领域) 会面临的挑战,在天文学界,即将上线的新型望远镜每天将会产生 Peta bytes 级的数据,要处理这么大数量的数据,需要新型的基础设施和先进的科学计算方法,而且要保证高效性和高速性。Szalay 通过阿姆达尔定律论述了平衡的观点。

(3) 从多个方面进行突破。要树立和推广一种普遍的方法来应对数据密集型科学计算的挑战。除了需要在计算硬件方面的投资,还需要在计算分析方法上进行投资研究。例如,数据收集方法分析结果常常不能够符合先前的假设,这些数据可能不是独立同分布的。这些现象对于收集来自于实验和物理系统观测站的数据来说是正常的。还有一个更为迫切的需求就是去开发确定性的,可扩展的分析算法,以及对几乎所有的硬件都支持的随机算法。一些情况下,数据量已经足够了,但有时科学家会面临一些语义方面的障碍,比方说在分析视频流信号的时候。

(4) 在过程中不断做出调整。还有一些情况就是,对于某些问题可能不能搜集到足够的数据,从而不能得到任何能站得住脚的结论。所以,要不断开发能够从有限的数据中提取信息的工具。随着可以分析使用的数据的增多,分析出来的结果可能也会各有不同,所以我们应该更加要在研究过程中的所有的阶段坚持科学的研究方法。获取更多更好的高质量数据肯定是必不可少的,但是数据本身是永远不能够代替繁重的分析工作的。

在信息的安全方面,国外的做法通常是设置安全机制,采用第三方信息安全审计,并对数据的使用作一些明确的规定,加大对信息窃取及修改的惩罚力度。

对企业组织者来说,首先需要盘点与己相关的数据资产,弄清楚哪些是自身拥有的数据,哪些是公共共享的数据,哪些是需要向第三方购买的数据,然后明确利用这些数据创造潜在价值,抓住其带来的机遇与挑战,同时从机构内部构造一个数据驱动型组织,制定相应的企业信息战略,最后再解决隐私和安全性方面的数据政策问题。对政策制定者来说,需要建立大数据有关的人力资源储备,通过激励机制促进数据共享,通过制定有关政策维持数据获利公司和其他利益单位之间的利益平衡,注重个人隐私,建立有效的知识产权保护体系,解决技术壁垒,确保对于基本信息和通信技术基础设施建设方面的投入。

5 "大数据"时代的赛博安全

新形势下的赛博安全,大数据处理系统的建立是必不可少的。在 2012 年 2、3 月份在美国旧金山召开的 RSA 安全会议上提出,大数据事件就是在网络安全行业中,推出新产品或者开创性的理论阐述的一个完美的舞台。RSA 主席 Art Coviello 已经讲

明，会增强自己在大数据的分析能力和组织武装自己的能力，来对抗日益增强的赛博威胁的冲击波。在当前，许多组织会搜集到大量安全方面的数据，但是，很多情况下，这些数据在安全层面上都是没用的，即使在有的时候是有用的，有些机构也会分析出错误的结论。目前大多数的系统仍然受限于控制误区和费时的更新中。安全措施基本上都是围绕常规审计和规范报告的，威胁识别几乎完全依赖于签名的恶意检测软件，这种单一的安全检测方式在日益增多的数据类型和数据形式的条件是远远不够的。在大数据时代下，这种状况必须要做出改变，使网络安全变得高效，机构单位需要在快如闪电的实时信息中筛选出威胁的存在。作为第一步，管理者需要监测网络中的每一部分，然后从所有可能的源头去搜集不同类型、不同格式的数据，从而来对攻击他们网络的所有的威胁有一个总体的概括性的了解 。当前的趋势是限制从安全控制中来搜集数据，大数据使机构能够检测不同新产品的动作方式之间的差异。已经搜集完数据之后，下一步是用高速分析的方式去关联这些搜集到的数据，产生一些操作性的信息，判别这些操作性信息的危险程度，以最快的速度作出反应，保护自己的网络数据不被侵犯，维持网络的正常运行。

6 美国在“大数据”时代针对网络安全做出的政策

“大数据”带来的网络安全问题，以及在宏观意义上的国家安全，不得不引起人们的重视。在今年 3 月份，奥巴马宣布“大数据的研究和发展计划”，通过提高从大型复杂的数字数据集中提取知识和观点的能力，帮助加快在科学与工程中的步伐，加强国家安全，改变教学研究。美国国防部先进研究项目局(DARPA) 为应对大数据时代的到来，宣布建立多个针对网络信息安全的研究项目。比较有代表性的有以下几个。

(1) 多尺度异常检测(ADAMS) 项目，该项目解决大规模数据集的异常检测和特征化。项目中对异常数据的检测指对现实世界环境中各种可操作的信息数据及线索的收集。最初的 ADAMS 应用程序只进行内部威胁检测，在日常网络活动环境中，检测单独的异常行动。

(2) 网络内部威胁(CINDER) 计划，旨在开发新的方法来检测军事计算机网络与网络间谍活动。作为一种揭露隐藏操作的手段，CINDER 将适用于将对不同类型对手的活动统一成“规范”的内部网络活动，并旨在提高对网络威胁检测的准确性和速度。

(3) Insight 计划，该计划主要解决目前情报、监视和侦察系统的不足，进行自动化和人机集成推理，使得能够提前对时间敏感的更大潜在威胁进行分析。该计划将会开发出资源管理系统，通过分析图像和非图像的传感器信息和其他来源的信息，进行网络威胁的自动识别和非常规的战争行为。

(4) 加密数据的编程计算(PROCEED) ，该研究工作旨在开发实用的方法，开发现代化计算编程语言，使数据加密时仍然能使用云计算环境，以克服信息安全的重大挑战。使用户能够不需要首次解密的情况下能够操纵加密的数据，它将使得对手拦截信息更加困难。

(5) 在视频和图像的检索和分析工具(VIRAT) 计划旨在开发一个系统能够利用军事图像分析员收集的数据进行大规模的军事图像分析。VIRAT 如果成功，将使分析师能够在相关活动发生时建立警报。VIRAT 还计划开发工具，能够以更高的准确率和召回率来从大量视频库里进行视频内容的检索。

(6) 任务导向的弹性云计划(Mission - oriented Resilient Clouds) 用来应对云计算固有的安全挑战，该项目要开发新的技术来检测攻击，并对攻击作出回应，高效地为云端建立起一个“区域健康体系”。项目的另一个目标是开发新技术使云端程序和设施能够在遭受赛博攻击的时候也能够完成相应的功能。在保证整个系统无大碍的情况下，个别主机或任务的损坏是可以容许的;

(7) XDATA 项目计划，旨在开发用于分析大量的半结构化和非结构化数据的计算技术和软件工具。最核心的挑战是，可伸缩的算法在分布式数据存储应用、如何使人机交互工具能够有效的迅速定制不同的任务，以方便对不同数据进行视觉化处理。对开源软件工具包的灵活使用，使得能够处理大量国防应用中的数据。

7 结 语

“大数据”时代的到来，充满了机遇与挑战，谁能够最快地习惯这种新形式下的数据模式，熟悉和掌握处理这种数据处理方法，谁就会在之后的信息战中占得先机，取得主动权。

本文在编写过程中，得到了信息产业部电子科技情报研究所乔榕高级工程师等专家的帮助，在此，表示感谢。